Research Design

Who Should Read This? You should read (study) the following essay if you are unfamiliar with any of the following terms.

- independent variable (explanatory or predictor or grouping variable)
- dependent variable (response or criterion variable)
- factor
- levels of a factor
- ANOVA (analysis of variance)
- experimental design
- true or randomized or designed experiment
- quasi-experimental design
- self-selected subjects
- intact groups
- observational studies
- confounding variables
- between subjects variables (designs)
- between groups designs (same as between subjects)
- within subjects variables (designs)
- repeated measures designs
- treatment by subjects designs
- matched groups (designs)
- simple designs
- single factor designs
- completely randomized designs
- balanced designs (vs. unbalanced designs)
- factorial designs
- mixed factorial designs
- main effects
- simple effects
- interaction effects
- correlation
- correlational designs
- regression analysis
- multiple regression
- analysis of covariance
- covariate

Part 1: Experimental Designs and ANOVA

Factors. All experiments involve trying to discover the effect of an *independent variable* on a *dependent variable*. The independent variable (IV) is the thing the investigator is manipulating, while the dependent variable (DV) is the thing the investigator is recording about his subjects. When

you read the description of an experiment, you will usually find that the investigator has divided his subjects into two or more groups. These groups are then treated differently in some fashion by the experimenter. You should ask yourself, "How are the different groups of subjects in this experiment treated differently?" When you answer that question, you will know the IV. The groups may be given different doses of a drug, for example. In this case, the IV would be drug dose.

In some cases, the investigator will not actually be manipulating the IV. The subjects will come into the "experiment" already selected for groups. This might occur, for example, if the investigator is looking for gender differences. Clearly in such a case, the male subjects will be considered one group in the study, and the female subjects will be considered another. Although the two groups are not "treated" differently by the investigator, they are still differentiated by gender. In this study, gender would be the "IV." Technically, such a study should not be considered a true experiment, which requires that subjects be randomly assigned to the treatment groups. ("True experiments" are also called *randomized experiments* for that reason.) When subjects are *self-selected*, the study is often referred to as a *quasi-experiment*, which literally means "something like an experiment but not quite exactly like an experiment." These studies are also often called *observational studies*, because we really haven't done anything experimental. We've merely observed which group our subjects are already in, and then we're recorded (observed) a value of the DV for each subject.

It rarely makes a difference to the statistical analysis, however, as long as certain conditions are met. Although quasi-experiments do not have true IVs--they have quasi-IVs--I will continue to refer to the "IV" in these cases. It might also be called the explanatory variable or the predictor variable. You should also not get too bent out of shape concerning the term *self-selected* subjects. True, people rarely get to choose their gender. What's important here is that the *investigator does not get to determine which subjects go into which group*. Such subjects are referred to as self-selected. The groups are called intact groups.

In the lingo of analysis of variance (ANOVA), independent variables are referred to as *factors*. That is to say, they are categorical variables that determine the grouping of the subjects, or which treatment condition the subjects are in. There may be more than one IV, or factor, in any given experiment or study, a complication we shall return to shortly.

A common mistake in talking about IVs is to name each value of the IV as an IV of its own. In the drug study, if subjects in one group are given 200 mg of caffeine, subjects in a second group are given 100 mg of caffeine, and subjects in a third group are given a placebo (0 mg of caffeine), it is a mistake to say that the IVs are 0 mg, 100 mg, and 200 mg of caffeine. Similarly, in a study investigating gender differences, it is a mistake to say that the IVs are men and women. In the first example, the IV is drug dose. In the second, it is gender. (I've even seen this mistake made in textbooks by people who should know better. It's a sure sign that the author doesn't know his knees from his elbows when it comes to experimental design!)

The individual values of an IV, or factor, are called *levels*. Thus, in the drug experiment, there is one IV (dose of caffeine), and it has three levels (0 mg, 100 mg, 200 mg). In the study on gender differences, there is one IV (gender), and it has the usual two levels (male, female), unless they've come up with a third since the last time I looked.

In the behavioral sciences, independent variables are often, if not usually, categorical variables, as in the case of the gender variable above. However, this is not always true. In the drug experiment,

illustrated above, the IV could also be considered numerical. We will discuss numerical IVs below, under correlational designs.

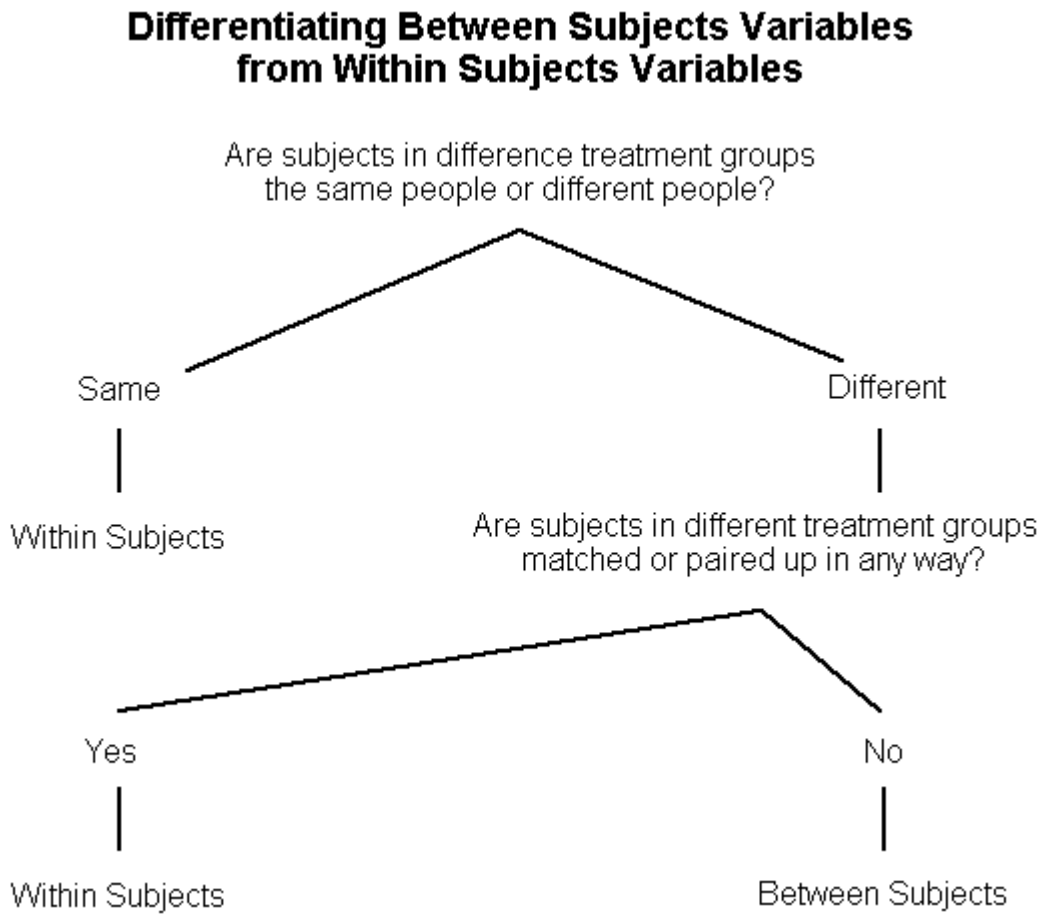
Confounding Variables. You should look very hard to find all the ways in which the treatment groups are differentiated from one another. Sometimes you'll find differences that were not at all obvious at first reading. Once you've made a list of these differences, you should ask yourself whether the investigator has established these differences intentionally because he or she wants to see if they change the values of the DV. In the studies above, the investigator is asking specific questions: Does drug dose matter? Does gender matter? Will changing the values of these variables change values of the DV? Once you've identified the questions being asked by the investigator, you've identified the IVs.

Sometimes, however, you'll come across differences between the treatment groups that the investigator seems to have overlooked. Or if he hasn't overlooked them, he is not interested in looking to see how they change values of the DV. These differences are called *confounding variables* or *confounds*. (To be entirely correct, they are only confounds if they actually *are associated with changes* in the values of the DV. Otherwise, they are merely potential confounds.) When confounds exist, then you do not know why the DV is changing. It could be the IV producing the change, which is what the investigator wants to see, but it could also be the confound. Statistically, it's often impossible to tell. Changes produced by a confound look just like changes produced by the IV to the statistical analysis. Sometimes confounds can be controlled for statistically, but this has to be planned in advance. When it's not, you've got a major problem interpreting the results of the study. Confounds are especially likely to occur when subjects are self-selected. The most important way of controlling for confounds is to *randomly assign subjects to treatment groups*. When this can't be done, or when it hasn't been done, watch out for confounds. You'll surely find them!

Between vs. Within Subjects Variables. Once you've identified how the different treatment groups in the study are being treated differently (or at least how they are different from one another in the case of self-selected subjects), the next question you should ask yourself is this. "Are the different treatment groups made up of different people (or animals or whatever the subjects are)?" If the answer to that question is "yes," then ask, "Are the subjects in the different groups in any way paired up or matched with each other?" If the answer to that question is "no," then you have a *between-subjects* variable, sometimes also called a *between-groups* variable. If the answer to the first question is "no," or if the answer to the first question is "yes" and the answer to the second question is "yes," then you have a *within-subjects* variable. The point is this:

- In an experiment with a between-subjects IV, the subjects in one group are being compared to different and entirely independent subjects in other groups.
- In an experiment with a within-subjects IV, the subjects in one group are being compared to themselves, or if not to themselves then to subjects like themselves with whom they have been matched or paired up.

Before you can begin a statistical analysis using ANOVA, you must know: (1) the identity of your IVs or factors, (2) which factors are being tested between-subjects and which are being tested within-subjects, and (3) what the DV is. Until you're sure you've identified these three things correctly, don't even think about reaching for your calculator or computer!



Simple Designs. Some experiments (and some quasi-experiments) incorporate only one IV or factor. The design of such an experiment is called a *simple design*, or sometimes a *single-factor* design. Suppose we wish to investigate the effect of caffeine on motor performance. We might proceed as follows. First, we would need to decide what aspect of motor performance we want to measure. That is, we need to decide what the DV is going to be. There are several things we could measure, but for the sake of illustration let's keep it as simple as possible. We'll measure the rate at which a subject can tap his fingertip on a button placed before him on a table. (We could also use a key on a computer keyboard, if we wanted to risk doing harm to the keyboard.) Second, we need to decide what doses of caffeine we are going to give (the levels of the IV). Let's go with placebo (none or 0 mg), a moderate dose (100 mg), and a high dose (200 mg). Furthermore, we decide to test 10 subjects at each dose. The basic procedure will be to have a subject come into the lab, take whichever dose of caffeine we have decided to test at this time, wait 20 minutes or so until the drug has had time to take effect, and then let the subject tap away for a minute at our keypad, which we have connected to an electronic counting device. (There is an important control we haven't mentioned here. Do you see it?) So far our design looks like this.

IV_1 = dose of caffeine		
0 mg	**100 mg**	**200 mg**
n = 10	*n* = 10	*n* = 10

We have one more important decision to make. Is the IV going to be tested between subjects or within subjects? The reasons for choosing one of these techniques over the other is a matter for your research methods course. Here I will simply describe the difference. Suppose we decide to use different people in the three groups. We have a subject pool of thirty people, and we randomly assign ten of them to each of the dosage-level groups. This would be called a *between-subjects design*. In this case, since we randomly assigned subjects to groups, it would also be called a *completely randomized design*. So we could thoroughly describe this experimental design by calling it a *single-factor* (or *simple*), *between subjects*, *completely randomized design.*

Alternatively, we could decide to use the same ten subjects over again in all three conditions of the experiment. Each subject would get, in turn, 0 mg, 100 mg, and 200 mg of caffeine, although not necessarily in that order, and hopefully separated by a few days to allow the previous treatment to wear off before the next one is begun. Here we would not be comparing the placebo group to the 100 mg group to the 200 mg group; we would be comparing Fred on placebo to Fred on 100 mg to Fred on 200 mg. This would be called a *single-factor within subjects design*. It could also be called a *repeated measures design*, which is roughly (although not quite exactly) the same thing as within subjects. And, in case you think we have gotten carried away with multiple names for things, I'll give you a third one. It is often also called a *treatment by subjects* design.

It's easy to slip into a within-subjects design without realizing it. Suppose we opt for the between-subjects design, but then we start to worry. Drug effects are influenced by body weight. Shouldn't we control for that? We notice that our subject pool contains Joe, Jeff, and John, all of whom have about the same body weight. So we decide to put Joe in one group, Jeff in another group, and John in the remaining group. We also match up other subjects by body weight and assort them to groups in a similar fashion. We now have a *matched groups design*. We are comparing Joe to Jeff to John. This is no longer between subjects but is now considered within subjects. (It is not repeated measures, however, because we are not measuring the same subjects repeatedly. It amounts to the same thing statistically though.) If you use the same subjects over again, or if you pair up or match your subjects in any way across the groups, your design is within subjects or repeated measures.

Balanced Designs. All of the examples we've discussed so far would be called *balanced designs*. That just means that all the treatment groups have the same number of subjects in them. Repeated measures designs have no choice but to be balanced, unless we forget to measure one of the subjects in one of the conditions, but between-subjects designs do not have to be balanced. We could easily put 11 subjects in one group, 10 in another, and 9 in the third, in which case we'd have an *unbalanced design*. In a randomized experiment, that would be a silly thing to do. (There are extenuating circumstances, but they are uncommon.) We pay a price for unbalanced designs in terms of our ability to detect the effect of the IV. If the design is factorial (next), then we also pay a price in the difficulty of the analysis. "Keep your designs balanced" is a good rule to follow. In the case of quasi-experimental designs, however, that is often not possible and may even be unwise if it jeopardizes random sampling.

Factorial Designs. It might well be the case that caffeine affects women differently from the way it affects men. To test for this, we decide to do our drug experiment on separate groups of men and women. Now, instead of having just one group of subjects getting each dose of drug, we have two groups, one group of men and one of women. Instead of having just three groups in our experiment,

we now have six groups. These groups are differentiated from one another on two separate dimensions: drug dose and gender. We have a group of men getting placebo treatment and a group of women getting placebo treatment. We have a group of men getting 100 mg and a group of women getting 100 mg. Finally, we have a group of men getting 200 mg and a group of women getting 200 mg. We might diagram this as follows. (Notice that design remains balanced.)

		IV_1 = dose of caffeine		
		0 mg	**100 mg**	**200 mg**
IV_2 = gender	**men**	*n* = 5	*n* = 5	*n* = 5
	women	*n* = 5	*n* = 5	*n* = 5

We now have two IVs, and they are crossed with one another in such a way that we are testing them both simultaneously, all levels of the first variable at all levels of the second variable. This is called a *factorial design*. The gender variable will naturally be tested between subjects. If we decide to do the same with the drug dose variable, then we have a *factorial design with two between-subjects factors*. Should we decide to test the drug dose IV within subjects, then we will have a special kind of factorial design called a *mixed factorial design*. The word *mixed* refers to the fact that we have some variables tested between subjects and some tested within subjects. In this case, we would say it is a *two-factor mixed factorial design with repeated measures on one factor*. (This is actually a little redundant, but it is better to be redundant than to leave something out.)

In either case, it would be called a 3x2 factorial design (read "three by two"), because we have three levels of the first variable crossed with two levels of the second variable. Such a design gives us 3x2=6 treatment conditions in the experiment.

The effect of caffeine might also change with repeated dosing, due to tolerance for example. To test for this effect, we might redesign the experiment as follows. First, we will omit the gender variable, and in its place we will put a days-of-treatment variable. Second, we have to decide how many levels this variable should have--let's say five. Of necessity, this is going to be a within-subjects variable. If we make dose-of-caffeine a between-subjects variable, then we will once again have a *two-factor mixed factorial design with repeated measures on one factor*. It's hard to imagine how we could test both of these variables within subjects, but suppose we find a way. In that event, we would have a *two-factor factorial design with two within-subjects factors*, or *repeated measures on both factors*. In either event, we have a 3x5 design that we would diagram this way.

		IV_1 = dose of caffeine		
		0 mg	**100 mg**	**200 mg**
IV_2 = days	**1**	*n* = 10	*n* = 10	*n* = 10
	2	*n* = 10	*n* = 10	*n* = 10
	3	*n* = 10	*n* = 10	*n* = 10
	4	*n* = 10	*n* = 10	*n* = 10
	5	*n* = 10	*n* = 10	*n* = 10

Gee, it's really a shame we had to leave out the gender variable. Well, actually we don't. We can incorporate gender as a third variable in a factorial design, making our design a three-factor design,

as 3x5x2 design. We might diagram it this way.

		IV_3 = gender					
		males			females		
		IV_1 = dose of caffeine			IV_1 = dose of caffeine		
		0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
IV_2 = days	1	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5
	2	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5
	3	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5
	4	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5
	5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5	*n* = 5

Now we have to decide which factors to make within subjects and which to make between subjects. To a certain extent, these decisions are made for us by nature. Gender will be between subjects. Subjects can only be tested repeatedly if we make days a within-subjects variable. So our only decision concerns dose. If we test dose between subjects, then we have a *three-factor mixed factorial design with repeated measures on one factor*. If we somehow find a way to test dose within subjects (it wouldn't be easy!), then we have a *three-factor mixed factorial design with repeated measures on two factors*. In principle, it's possible to have repeated measures on all factors, although that is clearly not possible in this particular experiment, unless we are planning some sex change operations as part of our design.

Main Effects. The change produced in the DV by an IV is called an *effect*. Technically, this definition of "effect" would apply only to experiments with a manipulated IV, which is to say with random assignment of subjects to groups and other appropriate controls for confounding variables, because it is only in these cases that we may assume a cause-and-effect relationship between the IV and DV. If the IV is not manipulated, for example because we don't have random assignment of subjects to treatment groups, then we have a quasi-experiment and the data are observational/correlational in nature. *The simple existence of correlation does not imply causality*.

Therefore, we will expand the definition of "effect" as follows: any change in the DV associated with a change in the IV is called an *effect*. Instead of saying "associated with," we could also say "related to" or "correlated with." We must be careful, however, when we have a quasi-experiment. If we go so far as to say a change in the DV was an "effect of the IV," then we have overstepped the limitations of observational data. This is a causal statement.

Nevertheless, it is convenient shorthand to refer to "an effect of an IV." Although we might prefer to say "this effect is associated with this IV," we often go for the lazy way out and say simply "this change in the DV is an effect of this IV." Each IV in an experiment has at least one potential effect associated with it. In a single factor experiment, if the value of the DV changes at different levels of the IV, then we have observed an "effect of the IV" (which may or may not be statistically significant and may or may not actually be caused by the IV). This type of effect is called a *main effect*. In factorial designs, there is a potential main effect associated with each of the IVs. Thus, in a two-factor design, we are looking for two potential main effects. In a three-factor design, we are looking for three potential main effects. And so on.

Suppose our summary statistics from the single-factor finger tapping experiment looked like this, where the means and standard deviations are in taps per minute. (One such experiment actually turned out this way.)

IV_1 = dose of caffeine		
0 mg	**100 mg**	**200 mg**
M = 244.8 sd = 2.394 n = 10	M = 246.4 sd = 2.066 n = 10	M = 248.3 sd = 2.214 n = 10

Notice that mean finger-tapping rate for the three groups differs. This would be described as a *main effect of dose of caffeine*. It's a main effect because the difference in the means is associated entirely with different levels of that one IV. It still remains to do a statistical test to see if the main effect is statistically significant. It is, as it turns out, and so we can say it is a *statistically significant main effect of dose of caffeine* [$F(2, 27) = 6.18$, $p = .0062$].

Now let's consider the experiment where we looked at both dose of caffeine and gender as our IVs, both tested between subjects. Let's suppose the results came out like this. (I made these up.)

		IV_1 = dose of caffeine		
		0 mg	**100 mg**	**200 mg**
IV_2 = gender	**men**	M = 243.6 sd = 2.608 n = 5	M = 246.0 sd = 2.915 n = 5	M = 249.6 sd = 1.673 n = 5
	women	M = 246.0 sd = 1.581 n = 5	M = 246.8 sd = 0.837 n = 5	M = 247.0 sd = 2.000 n = 5

To see if there is a main effect of dose, we need to calculate the overall means for dose, collapsing over gender. These are the same as in the previous table. So there is a main effect of dose of caffeine, and statistical analysis once again shows that it is statistically significant [$F(2, 24) = 7.28$, $p = .0034$]. To see if there is a main effect of gender we collapse over dose and calculate the overall means for men and women. We find the mean for men is $M = 246.4$ and for women is $M = 246.6$. It would be stretching a point to call this a main effect. Statistical analysis will undoubtedly show no significant difference here [$F(1, 24) = .07$, $p = .79$].

Simple Effects. *Simple effects* are like main effects, except that they are effects considered only at one level of other variables. For example, look at the means for men in the table above. Finger-tapping rate increases over doses of caffeine by 6.0 taps per minute, not a big increase perhaps, but a statistically significant one nevertheless [$F(2, 12) = 7.56$, $p = .0075$]. This is an example of a simple effect. If we look at the means for women, we see that they also increase over doses but not nearly as much. In fact, the change is not statistically significant [$F(2, 12) = .58$, $p = .57$]. Thus, there was no significant simple effect of caffeine dose on women.

Interaction Effects. When the simple effect of IV_1 is different at different levels of IV_2, then we have an *interaction* between IV_1 and IV_2. Here we have an interaction between dose of caffeine

and gender, and it is statistically significant, too. We would say we have *a statistically significant* dose x gender *interaction* [$F(2, 24) = 3.87$, $p = .035$]. The "x" symbol is read "by" when you are reading about the interaction. It is a "dose by gender" interaction. This would be called a *two-way* interaction, because it is an interaction between two variables. The key to recognizing an interaction effect is this: while it's possible to describe a main effect in words by referring to only one IV, it's impossible to completely describe an interaction in words without referring to at least two IVs. "Overall, finger-tapping rate increased with increasing doses of caffeine" (main effect). "However, with increasing doses of caffeine, finger-tapping rate increased more for men than it did for women" (interaction).

When a factorial design is used, there is a potential interaction effect for all possible combinations of the IVs. Let's consider the three-factor design described above, in which the IVs were dose, day, and gender. Each of these IVs has a potential main effect associated with it: a main effect of dose, a main effect of day, and a main effect of gender. There are also three potential two-way interactions: dose x day, dose x gender, and day x gender. Finally, there is also one potential three-way interaction: dose x day x gender.

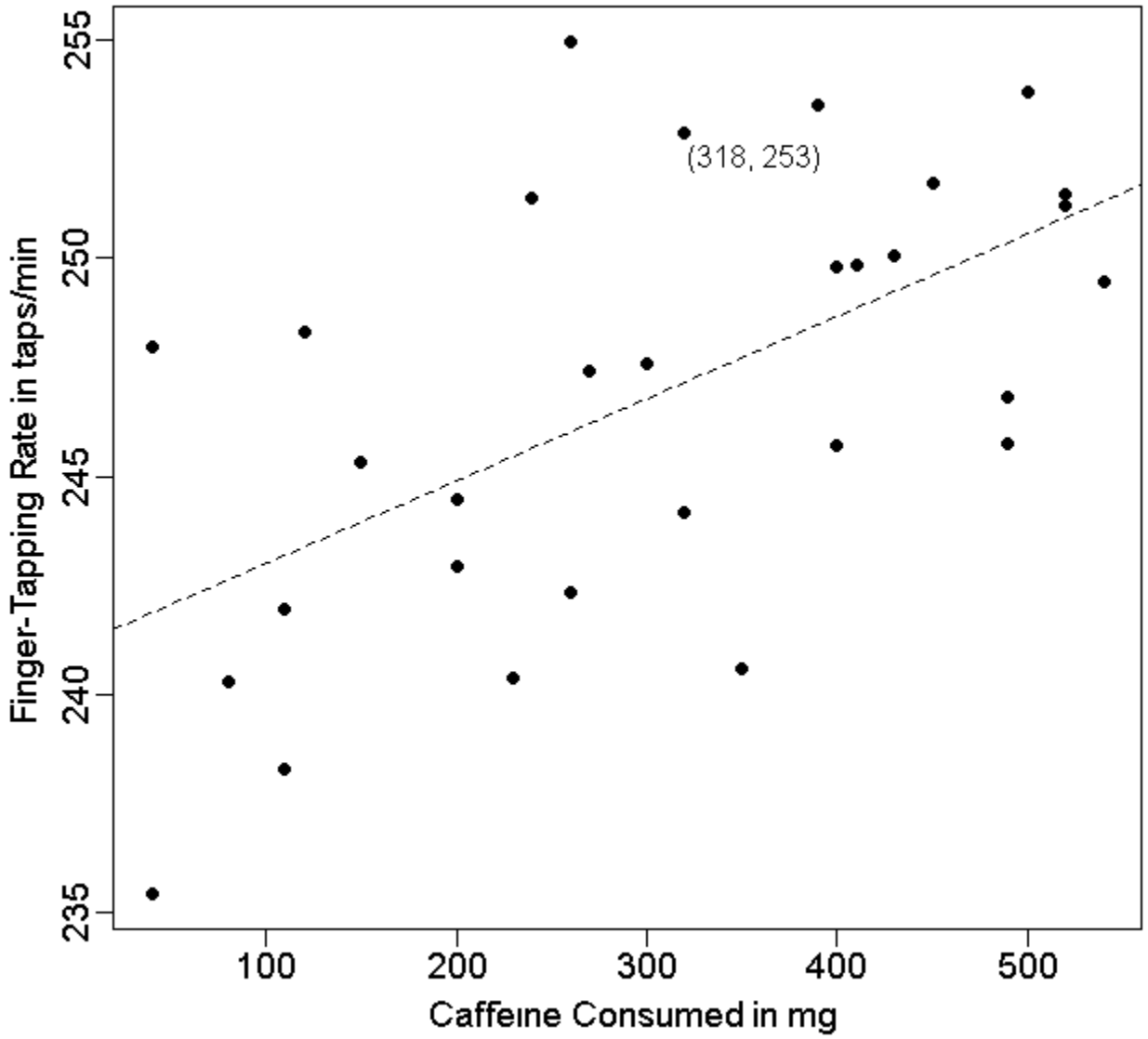
As the number of factors in an experiment increases, the number of potential interactions increases rapidly. With four IVs, there would be six potential two-way interactions, four potential three-way interactions, and one potential four-way interaction. Imagine trying to figure out what a four-way interaction looks like! Now imagine trying to explain it to your readers!!

Part 2: Correlational Designs and Regression

Designs With Numerical IVs. Let's continue with our attempt to find a relationship between amount of caffeine consumed and finger-tapping rate. One way we could do this is to have a group of people write down everything they eat for a 24-hour period and then come into the lab to have finger-tapping rate measured. We calculate the amount of caffeine in what each subject has eaten (in milligrams) and pair that value with the subject's finger-tapping score.

In this "experiment" we no longer have groups of subjects. Our IV now is numerical and varies continuously among our subjects (or potentially continuously). Thus, we can no longer test for a difference among group means. Instead, we look for a relationship between the number that represents the IV for each subject and the number that represents the DV. One subject's data might look like this: (318 mg of caffeine, 253 taps per minute).

This is called a *correlational design*. Correlational designs are also usually observational, because once again we have not manipulated a variable by randomly assigning anyone to conditions of an experiment. We would analyze the data from such a study by calculating a correlation coefficient or by plotting a scatterplot and determining the best fit regression line through the plotted points. Such a plot appears below.



The relationship between the two variables is shown not by a difference in means but by a pattern in the plotted points, namely that they trend from the lower left corner of the graph to the upper right corner. This indicates a positive correlation, and it is statistically significant ($r = .59$, $p < .001$). This is a simple correlational design (or a simple regression analysis) because there is one predictor variable (IV) and one response (DV), both of which are numerical.

Multiple Regression. Finger tapping rate also tends to slow down with age. (As does a lot of things!) So we might also want to record the age of our subjects and include that as a second numerical predictor in our regression analysis. We now have two numerical predictors (IVs) and a numerical response (DV). The design is still correlational (and observational), but the analysis would now be called a *multiple regression* (because of the multiple predictors).

In the last study, we elected to control for the age of the subjects by including age as a predictor in a multiple regression analysis. If there is an effect of age on finger tapping, the multiple regression should show it. More importantly (perhaps), if age and caffeine intake are confounded, the multiple

regression analysis should be able to tease apart that confound. For example, suppose younger people tend to drink much more caffeine than older people. Such a relationship between the IVs creates a confound, and multiple regression should allow us to tease these variables apart. ("Should" is not "will," however.)

Let's look at another possibility. Older people may respond the same to low amounts of caffeine as younger people do, but may respond less to higher doses. This is not a confound. This is an interaction effect. Multiple regression should allow us to see this as well.

Analysis of Covariance. Another way we could have controlled for the effect of age was to use subjects who are all the same age, or nearly the same age. We might use all college students, for example. We don't get to see an effect of age (if there is one) in the analysis, but at the same time, the DV is not "contaminated" by "nuisance variability" due to age that we may not be interested in. So let's adopt this strategy for our last study.

Let's say we have 30 subjects, just as in the very first experiment at the top of this discussion. Furthermore, we're not about to leave it to fate how much caffeine they consume. We're going to control caffeine dose experimentally. We're going to randomly assign 3 subjects each to 10 different doses of caffeine as follows: 0 mg, 25 mg, 50 mg, 100 mg, 150 mg, 200 mg, 250 mg, 300 mg, 400 mg, and 500 mg. The result would be a randomized experiment but with a numerical IV. There is nothing to stop us from analyzing this with ANOVA (as if the IV were any other grouping variable), but we would be silly to do so. The correct analysis would still be regression.

We're not going to let gender create any uninteresting nuisance variability because all of our subjects are going to be male. However, we are going to assign these males to two groups at random. One group will do the finger tapping test with their preferred hand, and the other group will do it with their nonpreferred hand.

We now have two IVs: dose of caffeine and hand. One we are treating as a numerical variable (dose), and the other is clearly a grouping variable. It looks like we have a peculiar sort of hybrid design that is a cross between the grouped designs we had initially and the correlational designs we've been discussing in Part 2. Can we handle this?

We can indeed! Different textbooks would call such a design different things, but it doesn't really matter much what we call it. I'm going to continue to call it correlational because of the existence of the numerical predictor (IV). And because we're going to analyze the data using regression analysis. Using what are called "dummy codes" for the grouping variable (hand), we can include it in the regression analysis just like any other numerical predictor. (Well, not *just* like!)

When we have a mixture of numerical predictors and factors as our IVs, we will call the regression analysis that we do an *analysis of covariance*. Not everyone agrees with this terminology. Analysis of covariance, or ANCOVA, can also be treated as an ANOVA, in which the numerical predictor is called the *covariate*. We'll discuss these issues when the time comes. All you need to know now is that such designs are possible.

I'll give you one more thing to think about, however. If we insisted on treating this last study as if we had two factors as IVs, is there any way we could have had a balanced design?